DOCUMENT RESUME

ED 236 699 CS 207 990

AUTHOR Shaw, Robert A.
TITLE Stability of Analytic Essay Scores: Implications for Diagnosis and Placement.
PUB DATE Aug 83
NOTE 21p.; Paper presented at the Annual Meeting of the American Psychological Association (91st, Anaheim, CA, August 26-30, 1983).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; Higher Education; *Placement; Research Methodology; Scoring; *Test Reliability; *Test Reviews; *Writing Evaluation; Writing Research

ABSTRACT

To study the stability of students' writing performances across occasions, four essays were obtained from each of 107 college students. These were scored with an analytic system that counts errors in 18 categories. Score reliability, computed by summing the 18 categories, was .36 for a single essay and .68 for the set of four essays. Score stability for students at the top and bottom of the score distribution was considerably higher than for students in the middle. As a research tool, the score reliability coefficient, in conjunction with counts of frequency, offers a measure of the relative complexity of particular writing skills. The implications for placement are clear. The reliability of a single essay, or even four essays, is lower than required for accurate placement into writing courses. (Tables of data are appended.) (Author/MM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Robert A. Shaw

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED236699

Stability of Analytic Essay Scores: Implications for Diagnosis and Placement

Robert A. Shaw

Brown University

Providence, RI 20912

Presented at the 91st Annual Convention of the American Psychological Association at Anaheim, California, August, 1983

CS 207990

Stability of Analytic Essay Scores

Robert A. Shaw

Abstract

A study of the score reliability of essays, i.e., the stability of a student's writing performance across occasions. Four essays were obtained from each of 107 college students. These were scored with an analytic system which counts errors in 18 categories. Score reliability of a combined score (computed by summing the 18 categories) was .36 for a single essay and .68 for the set of four essays.

Score stability for students at the top and bottom of the score distribution was considerably higher than for students in the middle. This finding is explained in terms of a domain of writing skills.

As a research tool, the score reliability coefficient, in conjunction with counts of frequency, offers a measure of the relative complexity of particular writing skills.

The implications for placement are clear. The reliability of a single essay, or even four essays, is lower than required for accurate placement into writing courses. However, the tests make fewer false positive placements (incorrectly identifying good writers as poor writers) than false negative (incorrectly identifying poor writers as good writers).

STABILITY OF ANALYTIC ESSAY SCORES: IMPLICATIONS FOR DIAGNOSIS AND PLACEMENT

The topic of score reliability has been sparsely represented in the literature on writing assessment. Virtually every book and article on the subject of writing assessment discusses reader reliability, but most ignore the equally important topic of score reliability, which is a measure of the stability of a student's writing performance from one occasion to another. Much of the fluctuation in writing scores from essay to essay is due not to reader error, but to changes in the quality of students' writing from one essay to another. This fluctuation can be considerable, and has serious implications for the placement of students in writing courses.

Over the last fifty years there have been a few well-crafted studies of the score reliability of essays. Three early studies assessed the score reliability of a single essay (Kincaid, 1953; Anderson, 1960; Traxler & Anderson, 1935). Two more recent studies determined the increased reliability obtained when more than one essay is scored (Breland & Gaynor, 1979; Godshalk, Swineford, & Coffman, 1966). In spite of a diversity of topics and scoring procedures, all five studies arrived at similar findings. A single essay, scored by two readers, was found to have a

score reliability of from .38 to .60; the score reliability of a set of five essays was found to be .84. The single-essay coefficients are lower than are generally considered acceptable for a test, and the five-essay coefficient of .84 is barely acceptable if the test is to be used for individual placement decisions.

The current study differed from previous studies in that it employed a quantitative, analytic scoring system, while previous studies have used holistic or qualitative scoring. Thus, the current study was able to yield detailed information about specific aspects of students' writing, facilitating an analysis of the mechanisms which cause fluctuations of writing quality from one essay to another.

Method

Subjects in the study were 107 students from a public junior college and two private universities. The students wrote four essays on assigned topics within a two-week period. Each was written in one 45-minute class period. The first and fourth topics were argumentative. Students were asked to agree or disagree with two statements: "learning is difficult, if not impossible, in American high

schools" and "anything that can be taught to another is relatively inconsequential." The second and third topics were descriptive: "describe a significant learning experience" and "describe an occasion in which you taught something to someone else."

The essays were scored using an analytic system which counts errors in eighteen categories (The system is described in detail in Van Nostrand, Pettigrew, & Shaw, 1980). Each category represents a linguistic or rhetorical constraint applying to essays, paragraphs, sentences, or words. Scoring consists of counting the number of times each constraint is misused. This raw error count is then divided by the number of essays, paragraphs, sentences, or words, as appropriate, to determine the category score. A total score was computed for each essay by summing the eighteen category scores. The categories are shown in Table 1.

Insert Table 1 about here.

Scoring was conducted by readers with extensive training in the use of the system. Each essay was read by a team of two readers, using an unusual but highly effective technique

called conference scoring. The first reader marked each error as he or she read. The second reader read each essay with the first reader's marks on it and noted all errors missed or improperly coded by the first reader. The two readers then discussed each essay and resolved any differences in the two readings of the essay.

In an earlier study using the same readers, the use of this conference scoring technique resulted in very high inter-rater reliability between two teams of readers, ranging from .93 to .99 in the eighteen categories (Horne, 1979). The high inter-reader reliability of this method of scoring means that virtually all the variations in the scores from essay to essay can be attributed to variations in the writing of students rather than to variations in marking by the readers.

Results

Table 2 shows score reliability estimates for a single essay and for the set of all four essays (All reliability coefficients were estimated with Cronbach's alpha). The estimated reliability of the total score was similar to that found in earlier studies: .36 for a single essay, .52 for

two essays, .62 for three essays, and .68 for the set of all four essays. Score reliabilities varied considerably among the eighteen categories. Reliability estimates for a single essay ranged from .02 to .56; estimates for the set of four essays ranged from .07 to .84.

Insert Table 2 about here.

To assess differences in score reliability among students of different ability, scores were divided into quartiles. A student was considered to be "stable" in a particular category if all four essays received marks in the same quartile or if three essays received marks in the same quartile and the fourth was in an adjacent quartile. The number of stable students in each quartile in each category is shown in Table 3. Quartile A contains students making the fewest errors; quartile D contains students making the most errors.

Insert Table 3 about here.

Table 3 adds valuable information to the numbers in Table 2. A large sub-set of students consistently scored in the same quartile on three or four essays. In the category scores, these students were generally those who made the fewest errors (quartile A). In the combined score, these students were those in the top (A) and bottom (D) quartiles. A total of 46 students, 39 of them in quartiles A and D, were stable on the combined score. These 39 stable students represent 72% of the 54 students who made up the top and bottom quartiles.

Discussion

The results of this study are best understood by referring to a domain of relatively independent writing skills. Any one essay samples only a small sub-set of the total domain of writing ability. The score reliability coefficient indicates the extent to which the sub-sets sampled by different essays overlap. This concept aids in the interpretation of the results of the study, particularly of differences in stability among students and differences in stability among the various categories in the scoring system.

Differences in Stability Among Students

The data in Table 3 indicate that the score reliability of a single essay is higher for students in the top and bottom quartiles of the distribution than for students in the middle quartiles. This finding can be explained with reference to the domain of writing skills. Consider students at three areas of the score distribution: high, low, and middle. Students of high ability were generally stable from essay to essay because they had mastered most of the constraints represented by the 18 categories, and they consistently used these writing constructions properly in all four essays. This stability is reflected both in their individual category scores and in their combined scores.

Students of low ability, on the other hand, had a poor grasp of most of the categories, and their category scores fluctuated from essay to essay, depending on the particular writing skills sampled in that essay. However, if the sampling was such that they avoided errors in one category, they made correspondingly more errors in another category. Thus, their combined score was consistently poor, even though their individual category scores varied considerably from essay to essay.

Students in the middle of the distribution had mastered some of the constraints but had trouble with others. Their

scores in the categories they had mastered remained stable from essay to essay, but their scores in the troublesome categories fluctuated considerably, driving their combined score up or down from essay to essay.

Only one previous study examined individual differences in score stability. Kincaid (1953) reported that, in general, the poorer writers in his sample showed more score stability from paper to paper than the better writers. His results seem in conflict with those of the current study. However, Kincaid determined the best and worst writers on the basis of their single highest essay score. Therefore, what he reported as fluctuation by the best writers was more likely fluctuation by students of average ability who, by chance, scored high on one essay. In short, Kincaid's results are consistent with those of the current study.

Differences in Stability Among Categories

The concept of the domain of writing skills also helps to explain why scores in some of the eighteen categories were more stable than others. The score reliability of an individual category is a function of the ratio between frequency of occurrence and size of the domain of that category. Categories which are sampled frequently in each essay relative to the size of the domain will have high

score reliability; categories which are sampled infrequently relative to the size of the domain will have low score reliability. For example, the category Joining Independent Clauses, which has a small but frequently-used domain consisting of three fairly simple rules, has a high score reliability, while other sentence categories, which represent larger domains of infrequently-used traits, have low score reliability. Conversely, the domain of Spelling is immense, but because the domain is sampled with every word the student writes, the score reliability of this category is high.

Data on frequency can be obtained simply by counting the occurrence of various constructions in a set of essays. Reliability coefficients, in conjunction with data on frequency of occurrence, can then offer valuable clues about the relative complexity of the traits represented in this or any other analytic scoring system.

Conclusion

The results of this study have implications for research on the writing process and for the use of analytic scores in the placement and diagnosis of students. As a research

tool, the score reliability coefficient, in conjunction with counts of frequency, offers a measure of the relative complexity of particular writing skills. This measure can be of great help in understanding why some skills are harder to learn than others. The ratio between frequency and reliability can easily be computed from the data provided by any quantitative scoring system.

The implications for placement are clear. The reliability coefficients for a single essay, or even four essays, are lower than required for accurate placement of students into writing courses. The results of the current study parallel those of earlier studies in this respect, even though the scoring system employed in this study was quite different from those employed in other studies. Regardless of the scoring system, the measured quality of a student's writing fluctuates considerably from one essay to another.

The figures in Table 3 have an important implication for the diagnosis of students' weaknesses within a writing course. Within individual categories of the scoring system, only students at the top of the distribution remained in the same quartile from essay to essay. The category scores of poor and average writers flucuated from essay to essay. On each essay, some category scores of some poor or average writers were in the top quartile. Thus, teachers who use

some type of error analysis to identify problems in the writing of individual students will make few false positive identifications (incorrectly identifying a good student as a poor student), but will make many false negative identifications (incorrectly identifying a poor student as a good student). The best remedy for this potential problem is to base diagnoses on several essays, though in some of the categories even the set of four essays had a score reliability of near zero.

Finally, in order to set these findings into context, it is worth noting that low reliability coefficients are not unique to measures of writing ability. In a massive review of the stability of a variety of human behaviors, Epstein (1979) reports that "With rare exceptions, studies that have correlated objective behavior on two occasions have obtained coefficients below .30" (p. 1097). Measures of writing ability are at least as reliable as, and generally more reliable, than the figure quoted by Epstein. His review, however, suggests that scores based on actual writing samples are unlikely to ever approach the reliability of scores based on multiple-choice tests. Nevertheless, assessments of writing samples are an indispensible tool in research on the writing process and in the teaching of writing. This study helps to clarify the limits within which such assessments can be used.

Note

The author acknowledges with thanks the assistance of Professors A. D. Van Nostrand and Joan Pettigrew in collecting the data and commenting on this manuscript.

References

Anderson, C. C. The new STEP essay test as a measure of composition ability. Educational and Psychological Measurement, 1960, 20, 95-102.

Breland, H. M. & Gaynor, J. L. A comparison of direct and indirect assessments of writing skill. Journal of Educational Measurement, 1979, 16, 119-126.

Cronbach, L. J.; Gleser, G. C.; Nanda, H.; & Rajaratnam, N. The dependability of behavioral measurements. New York: Wiley, 1972.

Coffman, W. E. On the validity of essay tests of achievement. Journal of Educational Measurement, 1966, 3, 151-156.

Epstein, S. The stability of behavior: I. On predicting most of the people much of the time. Personality and Social Psychology, 1979, 37, 1097-1126.

Godshalk, F. I., Swineford, F., & Coffman, W. E. The measurement of writing ability. New York: College Entrance Examination Board, 1966.

Horne, G. P. A report on statistical considerations in the selection and analysis of writing samples from Massachusetts State College students. Cambridge, Massachusetts: Commonwealth Center for High Technology/Education, January, 1979.

Kincaid, G. L. Some factors affecting variations in the quality of students' writing. (Doctoral dissertation, Michigan State University, 1953). Summarized in R. Braddock, R. Lloyd-Jones, & L. Schoer, Research in written composition. Champaign, Illinois: National Council of Teachers of English, 1963.

Traxler, A. E. & Anderson, H. A. The reliability of an essay test in English. School Review, 1935, 43, 534-539.

Van Nostrand, A. D., Pettigrew, J., & Shaw, R. A. Writing instruction in the elementary grades: Deriving a model by collaborative research. Final report, NIE-R-79-0010. Washington, D. C.: National Institute of Education, 1980. ERIC: ED 203 378.

Table 1

Analytic Scoring System Categories

Structural Unit	Category
Essay	no forecast
	lack of sequence between paragraphs
	no conclusion
Paragraph	no organizing idea
	missing evidence
	one-sentence paragraph
Sentence	fragment of an independent clause
	subject-verb disagreement
	verb form or tense misused
	inexact complement
	unparallel compounding
	wrong word choice
	diction (causing syntactic fracture)
	joining independent clauses unconventionally
	inaccurate punctuation demarcating phrases & clauses
	phrases & clauses misused
	pronoun referent not clear
Word	misspelling

Table 2

Score Reliability Estimates

For One and Four Essays

Category	One Essay	Four Essays
Forecast	.08	.27
Sequence	.19	.49
Conclusion	.28	.61
Organizing Idea	.22	.54
Evidence	.03	.10
One-Sentence Paragraph	.02	.10
Fragment	.18	.47
Subject-Verb	.02	.10
Verb	.30	.63
Complement	.06	.20
Compounding	.26	.59
Word Choice	.32	.65
Diction	.22	.52
Joining	.43	.75
Phrase & Clause Punctuation	.29	.62
Phrase & Clause	.11	.34
Pronoun Referent	.20	.50
Spelling	.56	.84
TOTAL SCORE	.36	.68

Table 3

Number of Stable Students,

By Quartile

	Quartile			
Category	A	B	C	D
Forecast	0	0	0	75
Sequence	42	0	0	4
Conclusion	5	0	0	45
Organizing Idea	8	0	2	6
Evidence	39	0	0	0
One-Sentence Paragraph	66	0	0	0
Fragment	59	0	0	0
Subject-Verb	69	0	0	0
Verb	24	0	1	6
Complement	55	0	0	1
Compounding	17	2	3	9
Word Choice	13	1	1	7
Diction	13	2	1	11
Joining	17	4	2	11
Phrase & Clause Punctuation	12	0	3	4
Phrase & Clause	22	0	0	4
Pronoun Referent	10	3	2	6
Spelling	18	8	9	13
TOTAL SCORE	20	5	2	19

Note to Table 3: A "stable" student was defined as one who received three scores in the same quartile and the fourth score in the same or an adjacent quartile. Numbers in some quartiles were greater than 27 (25% of 107) because some category distributions were skewed.